Breaking Bad: Extraction of Verb-Particle Constructions from a Parallel Subtitles Corpus

Aaron Smith
Department of Linguistics and Philology
Uppsala University
Box 635, 75126 Uppsala, Sweden
aaron.smith.4159@student.uu.se

Abstract

The automatic extraction of verb-particle constructions (VPCs) is of particular interest to the NLP community. Previous studies have shown that word alignment methods can be used with parallel corpora to successfully extract a range of multi-word expressions (MWEs). In this paper the technique is applied to a new type of corpus, made up of a collection of subtitles of movies and television series, which is parallel in English and Spanish. Building on previous research, it is shown that a precision level of $94 \pm 4.7\%$ can be achieved in English VPC extraction. This high level of precision is achieved despite the difficulties of aligning and tagging subtitles data. Moreover, many of the extracted VPCs are not present in online lexical resources, highlighting the benefits of using this unique corpus type, which contains a large number of slang and other informal expressions. An added benefit of using the word alignment process is that translations are also automatically extracted for each VPC. A precision rate of $75 \pm 8.5\%$ is found for the translations of English VPCs into Spanish. This study thus shows that VPCs are a particularly good subset of the MWE spectrum to attack using word alignment methods, and that subtitles data provide a range of interesting expressions that do not exist in other corpus types.

1 Introduction

In this paper, a method for the automatic extraction of English verb-particle constructions (VPCs) from parallel corpora is described and assessed. The method builds on previous research, particularly that of Caseli et al. (2010), adapting their approach specifically to VPC extraction and applying it to a different kind of corpus, based on subtitles from popular movies and television series, which is parallel in English and Spanish. The use of a parallel corpus also allows translations of VPCs to be obtained; an evaluation of the success rate of this process is also presented.

The paper is structured in the following manner: Section 2 discusses previous research and introduces key terminology, Section 3 describes the corpus and details the methodology and Section 4 explains the evaluation process. Results are then presented in Section 5, before discussion and future work in Section 6, and finally conclusions in Section 7.

2 Background

Amongst the many factors that contribute to the difficulty faced by NLP systems in processing multi-word expressions (MWEs), their sheer multifariousness is surely one of the most challenging. MWEs are combinations of simplex words that display idiosyncrasies in their syntax, semantics, or frequency (Caseli et al., 2010; Kim and Baldwin, 2010). They include nominal compounds such as *distance learning*, phrasal verbs such as *loosen up* and *rely on*, idioms such as *we'll cross that bridge when we come to it* and collocations such as *salt and pepper*, as well as instances which cannot so easily be classified such as *by the by* and *ad hoc* (Copestake et al., 2010). Due to their diverse and often non-compositional nature, MWEs constitute a big problem in many NLP tasks, from part-of-speech (PoS) tagging to parsing to machine translation (Chatterjee and Balyan, 2011, Constant et al., 2013).

In this paper the focus is on VPCs, a subset of phrasal verbs consisting of a verb and a particle, which, according to Villavicencio (2005), can be either prepositional, as in *hold on*, adverbial, as in *back away*, adjectival, as in *cut short*, or verbal, as

Proceedings of the 10th Workshop on Multiword Expressions (MWE 2014), pages 1–9,
Gothenburg, Sweden, 26-27 April 2014.

in *let be*. The definitions of phrasal verbs, VPCs and prepositional verbs are often confusing, with several competing terminologies. Greenbaum and Quirk (1990), for example, use a different system than that defined here: they use the term *multi-word verbs* where this study uses phrasal verbs, and *phrasal verbs* for those which are called VPCs here. In their system phrasal verbs are thus, along with prepositional verbs, a subset of multi-word verbs. The confusion between the different categories is often heightened by the fact that VPCs and prepositional verbs can be tricky to distinguish. The terminology used in this paper follows that of Villavicencio (2005): VPCs and prepositional verbs are a subset of the broader category of phrasal verbs.

The two most fundamental MWE-related tasks in NLP can be classified as *identification* and *extraction*. Identification, in the context of VPCs, is described in Kim and Baldwin (2010) as "the detection of individual VPC token instances in corpus data", while in extraction "the objective is to arrive at an inventory of VPCs types/lexical items based on analysis of token instances in corpus data". These tasks have relevance in different applications: identification is important in any form of text processing, whereas extraction is important for the creation of lexical resources and for text generation. Note that there is also a strong link between the two: lexical resources listing MWEs can naturally be used to identify their instances in a text.

In the present study the focus lies on VPC extraction: the goal is ultimately to create a list of valid VPCs. It is not the case that every verb can be combined with every possible particle – this would make our lives a lot easier (though perhaps less interesting). Villavicencio (2005) discusses the availability of VPCs in various lexical resources, including dictionaries, corpora, and the internet. She finds 3156 distinct VPCs across three electronic dictionaries, and extends that total to 9745 via automatic extraction from British National Corpus. She goes on to use the semantic classification of verbs defined by Levin (1993) to create lists of candidate VPCs based on their semantic properties, before using the internet as a gigantic corpus to attest them. The conclusion is that semantic classes are a good predictor of verbs' VPC productivity.

The current study owes a large debt to the work of Caseli et al. (2010). They proposed a method for identifying MWEs in bilingual corpora as a by-product of the word alignment process. Moreover, their method was able to extract possible translations for the MWEs in question, thus providing an efficient way to improve the coverage of bilingual lexical resources. Zarriess and Kuhn (2009) had previously argued that MWE patterns could be identified from one-to-many alignments in bilingual corpora in conjunction with syntactic filters. Caseli et al. (2010) draw on a previous study by Villada Moirón and Tiedemann (2006), who extract MWE candidates using association measures and head dependence heuristics before using alignment for ranking purposes.

An interesting variation on the word alignment extraction method was investigated by Liu (2011), who in fact use a monolingual corpus along with techniques designed for bilingual word alignment. They create a replica of the monolingual corpus, and align each sentence to its exact copy. They then adapt a word alignment algorithm (specifically IBM model 3), adding the constraint that a word cannot be aligned to its copy in the parallel corpus. This facilitates the extraction of collocations, and the authors show that their method elicits significant gains in both precision and recall over its competitors. A more recent attempt to use parallel corpora in the extraction of MWEs was made by Pichotta and DeNero (2013). They focused on English phrasal verbs, and devised a method of combining information from translations into many languages. They conclude that using information from multiple languages provides the most effective overall system.

A key finding of Caseli et al. (2010) was that their method achieved its highest levels of precision for phrasal verbs. For this reason the present study will focus specifically on VPCs, in a sense narrowing the previous study to focus on part of its most successful element. Like that study, this work will also find and evaluate candidate translations for each extracted English phrase. The corpus used in that study was composed of articles from a Brazilian scientific magazine. Based on the observation that VPCs are often less formal than their non-VPC counterparts (consider for example *The experiments back up the theory* v. *The experiments support the theory*), the current work evaluates the methodology on a spoken text corpus, specifically subtitles from movies and televi-

sion series. It is expected that this type of corpus will have a high density of VPCs, and moreover that they will often be informal, slang, and even profanities that would not be found in most corpus types. Indeed, the name of one of the most successful television series of recent times, *Breaking Bad*, is a perfect example of a slang VPC that would not be found in most lexical resources.

3 Methodology

The methodology in this study, adapted from that of Caseli et al. (2010), consists of four stages: PoS tagging, extraction, filtering and grouping, which are explained in turn in Sections 3.1–3.4. The corpus used is the OpenSubtitles2012 corpus (Tiedemann, 2012), a collection of documents from http://www.opensubtitles.org/, consisting of subtitles from movies and television series. As it based on user uploads there can be several sets of subtitles for the same movie, normally varying only slightly from each other. The corpus is tokenised, true-cased and sentence-aligned, and various word alignments are also provided. The section of the corpus used in this study, which is parallel in English and Spanish, contains 39,826,013 sentence pairs, with 342,833,112 English tokens and 299,880,802 Spanish tokens.

3.1 PoS Tagging

First of all, both the English and Spanish data are PoS tagged using TreeTagger (Schmid, 1994). An advantage of TreeTagger is that as well as PoS tags, it also provides lemma information for each word, which will be useful later in identifying different conjugations of the same VPCs. Subtitles, being a form of spoken text, are inherently difficult to tag; the overall accuracy of the TreeTagger is likely to be low on this data type. It should be noted however that PoS taggers generally have a high accuracy for verbs compared to other parts of speech.

3.2 Extraction

Using the `aligned.grow-diag-final-and` alignment file provided with the corpus, all word alignments containing more than one word in either language are extracted. This alignment file has been created by first word-aligning the parallel data sets in both directions using GIZA++ (Och and Ney, 2000), before merging them according to the algorithm in Och and Ney (2003). By varying the parameters to this algorithm to trade between precision and recall, various other alignment files have also been produced and made available as part of the OpenSubtitles2012 corpus.

The first alignment from the raw extraction process (for illustration purposes – there is nothing particularly special about this entry) is as follows:

```
've/VHP/have got/VVN/get ///
 tengo/VLfin/tener
```

The English *'ve got* is aligned to the Spanish *tengo* ("I have"), along with the respective PoS tags and lemmas. In total there are 53,633,153 such alignments in the corpus, many of which are repetitions. Identical entries are counted and sorted, before filtering is applied to find candidate VPCs.

3.3 Filtering

This is achieved by looking for all instances where the first English word has a verb tag (any tag beginning with *V*), the second is a particle (indicated by the tag *RP*), and the Spanish translation is also a verb. A minimum frequency of five is also effected; this is higher than the threshold of two applied by Caseli et al. (2010). There are several reasons for this: the larger corpus size here, the fact that PoS tagging is expected to be less accurate on this corpus, and the fact that some movies have more than one set of subtitles, leading to some almost identical sections in the corpus. This filtering is rather strict: to make it through this stage a VPC must occur at least five times in the corpus in exactly the same conjugation with the same translation. Some genuine VPCs might therefore be filtered away at this stage; those that occur few times and in different conjugations will be lost. The value of five was chosen early on in the study and left unchanged, based on some initial observations of lines that were repeated two or three times in the corpus and taking into account the other factors mentioned above. This parameter can of course be adjusted to increase recall, with the expected damage to the precision score; a more detailed investigation of this effect would be an interesting extension to the present study.

The filtered list contains a total of 18186 entries, the first of which is:

```
10900 come/VV/come on/RP/on ///
 vamos/VLfin/ir
```

This looks promising so far: the English entry *come on* is a valid VPC, and the Spanish translation *vamos* ("let's go") is a good translation. There

is still more work to do, however, as at this stage the list contains many instances of the same VPCs in different conjugations and with different translations. There are also, due to the fact that the original corpus was in true case, some instances of repetitions of the same VPC with different casing.

3.4 Grouping

The remaining data is lower-cased, before entries are grouped based on their lemmas, adding together the respective counts. By doing this some information is lost: certain VPCs may only naturally appear in certain conjugations, or may have different meanings depending on the conjugation they appear in. This therefore undoubtedly introduces some error into the evaluation process, but for the purposes of simplification of analysis is a crucial step.

Grouping reduces the list of VPC-translation pairs to 6833 entries, 37.6% of the number before grouping. This large reduction shows that the VPCs that occur many times in one conjugation tend to also appear in several other conjugations. The grouping process merges these to a single entry, leading to the observed reduction. Amongst the remaining 6833 entries, there are 1424 unique English VPCs. The next challenge is to evaluate the accuracy of the results.

4 Evaluation

The evaluation of the extracted candidate VPCs and their translations is in three parts: first, an evaluation of whether the candidates are in fact valid English VPCs; secondly, whether they already exist in certain online resources; and thirdly whether the Spanish translations are valid. Evaluating all 6833 candidates is not feasible in the time-frame of this study, thus the following approach is taken: a random selection of 100 VPC candidates is chosen from the list of 1424 VPCs, then for each of these candidates the highest probability translation (that with the highest count in the corpus) is found.

4.1 Validity of VPC Candidates

The 100 candidate VPCs are judged by a native English speaker as either valid or not, following the definitions and rules set out in Chapter 16 of Greenbaum and Quirk (1990) (note however their different terminology as mentioned in Section 2). One of the major difficulties in this evaluation is that VPCs are productive; it can be difficult even for a native speaker to judge the validity of a VPC candidate. Consider for example the unusual VPC *ambulance off*; while this almost certainly would not appear in any lexical resources, nor would have been uttered or heard by the vast majority, native speaker intuition says that it could be used as a VPC in the sense of 'carry away in an ambulance'. This should therefore be judged valid in the evaluation. It is important to remember here that one of the main reasons for using the subtitles corpus in the first place is to find unusual VPCs not usually found in other corpora types or lexical resources; candidates cannot simply be ruled out because they have never been seen or heard before by the person doing the evaluation. *Ambulance off* does actually appear in the corpus, in the sentence *A few of a certain Billy-boy's friends were ambulanced off*, though it is not part of the 100 candidate VPCs evaluated in this study.

At the evaluation stage, the aim is to judge whether the candidate VPCs could in theory validly be employed as VPCs, not to judge whether they were in fact used as VPCs in the corpus. The corpus itself was however a useful resource for the judge; if a borderline VPC candidate was clearly used at least once as a VPC in the corpus, then it was judged valid. Not all VPC candidates were checked against the corpus however, as many could be judged valid without this step. It is worth noting that some genuine VPCs could have found themselves on the candidate list despite not actually having been employed as VPCs in the corpus, though this probably happens very infrequently.

4.2 Existence in Current Lexical Resources

Once valid VPCs have been identified by the judge from the list of 100 candidates in the previous step, they are checked against two online resources: Dictionary.com (http://dictionary.reference.com/) and The Free Dictionary (http://www.thefreedictionary.com/). Both these resources contain substantial quantities of MWEs; The Free Dictionary even has its own 'idioms' section containing many slang expressions. A VPC is considered to be already documented if it appears anywhere in either of the two dictionaries.

4.3 Accuracy of Translations

The final stage of evaluation was carried out by a native Spanish speaker judge from Mexico with a near-native level of English. The judge was asked to asses whether each of the Spanish translation candidates could be employed as a translation of the English VPC in question. The original corpus was used for reference purposes in a similar manner to the evaluation of the VPC candidates: not every example was looked up but in borderline cases it served as a useful reference.

5 Results

5.1 Validity of VPC Candidates

Amongst the 100 randomly selected VPC candidates, 94 were judged valid by a native speaker. The normal approximation gives a 95% confidence interval of $94 \pm 4.7\%$. In the original list of 1424 candidates, the number of true VPCs is therefore expected to lie in the range between 1272 and 1405. This precision rate is in line with the figure of 88.94–97.30% stated in Table 9 of Caseli et al. (2010). Note however that the two figures are not directly comparable; in their study they looked at all combinations of verbs with particles *or* prepositions, and judged whether they were true MWEs. Their analysis thus likely includes many prepositional verbs as well as VPCs. Remember here that only combinations of verbs with particles were considered, and it was judged whether they were true VPCs. The current study shows however that high levels of precision can be achieved in the extraction of phrasal verbs, even given a more difficult corpus type.

Amongst the VPC candidates judged valid, four appeared in slightly unusual form in the list: *teared up*, *brung down*, *fessed up* and *writ down*. In all four cases the problem seems to stem from the lemmatiser: it fails to convert the past tense *teared* to the infinitive *tear* (note that "tear" has two quite separate meanings with corresponding pronunciations – one with "teared" as past tense and one with "tore"), it fails to recognise the dialectal variation *brung* (instead of *brought*), it fails to recognise the slang verb *fess* (meaning "confess"), and it fails to recognise an old variation on the past tense of *write*, which was *writ* rather than *wrote*. These mistakes of the lemmatiser are not punished; there were marked valid as long as they were genuine VPCs. This reinforces a difficulty of working with subtitle corpora: verbs might be used in unusual forms which cause difficulties for existing automatic text-analysis tools. It is of course also the reason why subtitles are in fact so interesting as corpus material.

It is illuminating to analyse why certain VPC candidates were judged invalid; this can highlight problems with the method, the evaluation, or even the corpus, which may help future studies. The six VPC candidates in question are *base on, bolt out, bowl off, bury out, hide down* and *imprint on*. These false positives all contain valid verbs, but combined with the particle do not make valid VPCs. In several cases the confusion arises between a preposition and a particle; it appears the tagger has incorrectly labelled the second token as a particle instead of a preposition in the cases *base on*, *bolt out*, *bury out* and *imprint on*. This seems to occur particularly when the preposition occurs at the very end of a sentence, for example in *that's what these prices are based on*, or when there is a two-word preposition such as in phrases like *he bolted out of the room*. It is easy to see how the tagger could have interpreted these prepositions as particles; very similar examples can be found where we do indeed have a VPC, such as *that was a real mess up* or *he was shut out of the discussion* (the particles 'up' and 'out' here appear in the same positions as the prepositions in the previous examples). The candidate VPC *hide down* is a somewhat similar case, appearing in phrases such as *let's hide down there*. The tagger incorrectly labels 'down' as a particle instead of an adverb. A clue that this is the wrong interpretation comes from the fact that when the phrase is spoken out loud the emphasis is placed on *hide*.

The final false positive to be explained is *bowl off*. This verb appears in the phrase *they'd bowl you off a cliff*, which occurs no less than eleven times in the corpus, each time aligned to a single Spanish verb. Here we see how a problem with the corpus leads to errors in the final list of candidates. This appears to be a case where several sets of subtitles exist for the same movie, and the tagger and aligner are making the same faulty decision each time they see this phrase, allowing the incorrect VPC to bypass the filters. One possible resolution to this problem could be to simply exclude all identical lines above a certain length from the corpus. This is however somewhat unsatisfactory, as having multiple copies of the same subtitles does provide some information; the fact that

several users have all chosen to transcribe a particular section of a movie in a certain way should increase our credence in the fact that it is both valid English and an accurate reflection of what was actually said. Another option might therefore be to alter the parameter determining the minimum number of times a particular alignment must occur to be included in the analysis. A more thorough investigation of the trade off between precision and recall, which can be altered both by varying this parameter and by invoking more or less strict word alignment algorithms, could be the subject of a further study.

It is reasonable to ask the question as to why the accuracy of VPC extraction is so high in comparison to other MWE types. A possible reason for this is that VPCs in one language, such as English, tend to be translated to a verb construction in another language, such as Spanish. They can thus said to be cross-linguistically consistent (although not in the stronger sense that a VPC always translates to a VPC – many languages indeed do not have VPCs). This is not true of all MWE types; in many cases complex constructions may be required to translate a certain type of MWE from one language to another. Another contributing factor may be that PoS taggers have good accuracy for verbs compared to other PoS categories, which makes the filtering process more precise.

5.2 Existence in Current Lexical Resources

One of the aims of this study was to show that subtitles data contain interesting VPCs that are rarely seen in other types of corpora, even those that contain a considerable number of idioms and slang expressions. Of the 94 validated VPCs from Section 5.1, 80 were found on either Dictionary.com or The Free Dictionary. 14 of the 100 randomly selected VPC candidates were thus valid previously undocumented VPCs (see Table 1), with a 95% confidence interval of $14 \pm 6.8\%$. This gives us a range of valid previously undocumented VPCs amongst the total 1424 extracted between 103 and 296.

beam up	make whole
clamber up	reach over
dance around	shorten up
grab up	single up
grill up	spin up
lift up	storm off
poke up	torch out

Table 1: The 14 validated VPCs that do not appear in either of the online resources.

Interestingly, nine of the 14 previously undocumented VPCs in the sample take the particle 'up', suggesting that this type of VPC may be particularly under-represented in lexical resources. This particle often adds an aspectual meaning to the verb in question, rather than creating a completely new idiomatic sense. That is certainly the case with several of the VPCs listed in Table 1; *shorten up*, *grab up* and *grill up*, for example, could be replaced by *shorten*, *grab* and *grill* respectively without a dramatic change in sense. This particle may therefore be somewhat more productive than the others observed in Table 1; *whole*, *out*, *over*, *around*, and *off* cannot be so freely added to verbs to make new VPCs.

5.3 Accuracy of Translations

The translations of 75 of the 94 validated VPCs from Section 5.1 were judged valid by a native Spanish speaker. This equates to a 95% confidence interval of $75 \pm 8.5\%$ of the original selection of 100 VPC candidates that are valid and have correct translations. As with the original list of English VPCs, there were some issues in the Spanish translations stemming from the lemmatiser. Certain verbs appeared in forms other than the infinitive; as before these mistakes were not punished in the evaluation. The point here was not to judge the quality of the lemmatisation, which was primarily used as a tool to simplify the evaluation.

The precision rate of $75 \pm 8.5\%$ obtained in this study is higher than the range 58.61–66.91% quoted in Caseli et al. (2010), though there is a small overlap of 0.41% (note that their range is bounded by the number of examples judged correct by two judges and those judged correct by only one of the judges, and is not a statistical confidence interval in the same sense). Their analysis again differs somewhat here, however, as they consider translations of many different types of MWE; they do not present an analysis of how this figure breaks down with different MWE types. The results presented here suggest that high precision rates can be achieved for VPC translations using this alignment method. Although the precision is a little lower than for VPC extraction, it is still likely to be practically quite useful in the creation of bilingual lexical resources for NLP tasks.

6 Discussion and Future Work

The methodology described in this paper consisted of four stages – PoS tagging, extraction, filtering and grouping. Analysis of false positive candidate VPCs extracted from the corpus demonstrated that improvements at various points along this pipeline could be effected to boost the final results. A common error at the first stage was prepositions being tagged as particles. It was always likely that PoS tagging on difficult data like subtitles would be less than perfect, and for this reason it is not surprising that errors of this nature arose. Training a PoS-tagger on labelled subtitles data, something which is not currently available, would be an obvious way to improve the accuracy here.

An important factor at the extraction stage was that some sections of the corpus were essentially duplicates of each other, due to the fact that there could be several user uploads of the same movie. This could lead to certain VPCs being validated despite being very rare in reality. A solution here might be to try to remove duplicates from the corpus, and there are several conceivable ways of doing this. One could impose a limit of one set of subtitles per movie, though this would require access to a version of the corpus with more information than that used in this study, and would raise the question of which version to choose, bearing in mind that both the English and Spanish subtitles may have several versions. A more brute method would be to directly remove duplicate lines from the corpus, that is to say all lines where both the English and Spanish are identical in every respect. A preliminary study (not shown here) shows that keeping all other parameters equal, this reduces the number of candidate VPC-translation pairs from 6833 to 3766 (a reduction of 45%), with a reduction in the number of unique VPCs from 1424 to 852 (a reduction of 40%). One would of course hope that the precision rate be higher amongst the candidate VPCs, though given the large reduction of candidates, the overall number of valid VPCs extracted would surely be lower. A lowering of the frequency threshold might therefore be required in order to extract more VPCs; a future study will look into this trade-off.

Another methodological choice made in this study was the order in which various parts of the methodology were carried out: grouping came after filtering in the four-stage process, but these could equally be switched. A preliminary study (not shown here) shows that applying the grouping algorithm before the frequency threshold increases the number of candidate VPCs to 12,945 (an increase of 89%), with 2052 unique VPCs (an increase of 44%). However, there is a corresponding decrease in precision from $94 \pm 4.7\%$ to $85 \pm 7.0\%$ (though the confidence intervals do overlap here). A more thorough investigation would be required to confirm this effect, and to test what happens to the number of previously undocumented VPCs and precision of translations.

The frequency threshold was set to five in this work: each candidate VPC had to appear at least five times in the same conjugation to be accepted. This number was chosen at the beginning of the study and never altered; it is clear however that it plays a big role in the final number of candidate VPCs and the precision rate therein. An interesting extension to this work would be to analyse the relationship between this threshold and precision: at what frequency level does the precision become acceptable? This could be analysed from both the point of view of VPC candidates and their translations: the level may not be the same for both. This would of course require a large amount of empirical evaluation that may be expensive and hard to carry out in practise. The highest frequency translations for each of the randomly selected VPC candidates were evaluated in this study; it would also be interesting to look at the precision rate for all translations. Caseli et al. (2010) found that the range of accurate translations reduced from 58.61–66.92% for the most frequent translations to 46.08–54.87% for all possible translations across a larger spectrum of MWEs.

The results presented in this study would be stronger if confirmed by other judges; the more the better but ideally at least three. It should be remembered however that the criteria for judging was whether the VPC candidate could in any circumstance be used as a genuine VPC. Only one positive example is required to prove this for each VPC candidate, and no number of negative examples proves the reverse. The difficulty for the judge is therefore not really that he or she will accidentally label an invalid candidate as valid, but the opposite: sometimes it is simply difficult to think up a valid phrase with the VPC in question, but once it appears in the mind of the judge he is certain that it is valid. The same can be true of translation: it may be difficult to think of a sense

of the English VPC in which the Spanish verb is valid, even if that sense does exist. The results presented here can thus be viewed as a minimum: the addition of further judges is unlikely to lead to a reduction in precision, but could lead to an increase. One area where further evaluation could lead to less-impressive results is the number of undocumented VPCs. Validated VPCs were checked against two resources in this study: The Free Dictionary and Dictionary.com. It would be interesting to do further tests against other resources, such as the English Resource Grammar and Lexicon (www.delph-in.net/erg/).

This study did not consider recall, choosing instead to focus on precision and a comparison of extracted VPCs with existing resources. It would however be useful for many applications to have an idea of the percentage of VPCs in the corpus that end up in the final list, although a full analysis would require a labelled subtitles corpus. Caseli et al. (2010) present a method to estimate recall when a labelled corpus is not available. Generally speaking however it can be assumed that the normal inverse relation between precision and recall holds here. The exact dynamic of this relation can be adjusted in the filtering process: by letting VPCs with lower frequency through recall is bound to increase, but at the same time reduce the high levels of precision as more false positives end up in the final list. The balance between precision and recall can also be adjusted during the alignment process; the effect this would have on VPC extraction is unclear. An evaluation of this effect could be carried out by re-running the study using each of the different alignment tables provided with the OpenSubtitles corpus.

Only one language pair was considered in this study, namely English and Spanish. Pichotta and DeNero (2013) have shown that combining information from many languages – albeit in conjunction with a different extraction method – can improve VPC extraction accuracy. One way to further increase the precision achieved via the alignment methods in this study may be to use a similar combination technique. The latest version of the OpenSubtitles corpus contains 59 different languages, and this multitude of data could potentially be put to better use to obtain yet more VPCs. The choice of English and Spanish is also relevant via the fact that English has VPCs while Spanish does not – this may be an important factor. Whether better results could be obtained using two languages with VPCs, such as English and German, for example, is another interesting question that may be the subject of a follow up study.

7 Conclusions

This study has demonstrated that word alignment methods and a PoS tag based filter on a large parallel subtitles corpus can be used to achieve high precision extraction of VPCs and their translations. Despite the difficulties associated with the corpus type, which hinder both the tagging and the word alignment processes, a precision of $94 \pm 4.7\%$ was found for the extraction of valid English VPCs from a parallel corpus in English and Spanish. $14 \pm 6.8\%$ of the extracted VPC candidates were both valid and previously undocumented in two large online resources, while several more appeared in unusual dialectal forms, highlighting the unique nature of the corpus type. Analysing the Spanish translations extracted along with the VPCs, $75 \pm 8.5\%$ were judged valid by a native Spanish speaker. This represents a large increase in precision over similar previous studies, highlighting the benefits of focusing on VPCs rather than a larger range of MWE types.

Acknowledgements

This work benefited greatly from discussions with my fellow students on the Language Technology: Research and Development course at Uppsala University. I am particularly grateful to Nina Schottmüller and Marie Dubremetz for their detailed suggestions, and our teacher Joakim Nivre for his significant input to this paper. I would also like to thank the three anonymous reviewers for their valuable feedback.

References

H. M. Caseli, C. Ramisch, M. G. V. Nunes, and A. Villavicencio. 2010. Alignment-based extraction of multiword expressions. *Language Resources & Evaluation*, 44:59–77.

N. Chatterjee and R. Balyan. 2011. Context Resolution of Verb Particle Constructions for English to Hindi Translation. *25th Pacific Asia Conference on Language, Information and Computation*, 140–149.

M. Constant and J. Le Roux and A. Signone. 2013. Combining Compound Recognition and PCFG-LA

Parsing with Word Lattices and Conditional Random Fields. In *ACM Transactions on Speech and Language Processing*, 10(3).

A. Copestake, F. Lambeau, A. Villavicencio, F. Bond, T. Baldwin, I. Sag, and D. Flickinger. 2002. Multiword expressions: linguistic precision and reusability. In *Proceedings of LREC*, 1941–1947.

C. M. Darwin and L. S. Gray. 1999. Going After the Phrasal Verb: An Alternative Approach to Classification. *TESOL Quarterly*, 33(1).

S. Greenbaum and R. Quirk. 1990. *A Student's Grammar of the English Language*. Pearson Education Limited, Harlow, UK.

S. N. Kim and T. Baldwin. 2010. How to pick out token instances of English verb-particle constructions. *Language Resources & Evaluation*, 44:97–113.

B. Levin. 1993. *English Verb Classes and Alternations – A Preliminary Investigation*. The Chicago Press.

Z. Liu, H. Wang, H. Wu, and S. Li. 2011. Two-Word Collocation Extraction Using Monolingual Word Alignment Method. In *ACM Transactions on Intelligent Systems and Technology*, 3(487–495).

F. J. Och and H. Ney. 2000. Improved Statistical Alignment Models. In *Proceedings of the 38th Annual Meeting of the ACL*, 440–447.

F. J. Och and H. Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(19–51).

K. Pichotta and J. DeNero. 2013. Identifying Phrasal Verbs Using Many Bilingual Corpora. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 636–646.

H. Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

J. Tiedemann. 2012. Parallel Data, Tools, and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, 2214–2218.

B. Villada Moirón and J. Tiedemann. 2006. Identifying Idiomatic Expressions using Automatic Word-Alignment. In *Proceedings of the Workshop on Multi-Word-Expressions in a Multilingual Context (EACL-2006)*, 33–40.

A. Villavicencio. 2005. The availability of verb particle constructions in lexical resources: How much is enough? *Computer Speech And Language*, 19:415–432.

S. Zarriess and J. Kuhn. 2009. Exploiting Translational Correspondences for Pattern-Independent MWE Identication. In *Proceedings of the Workshop on Multiword Expressions*, Suntec, Singapore 23–30